AAPM: Large Language Model Agent-based Asset Pricing Models

Anonymous EMNLP submission

Abstract

In this study, we propose a novel asset pricing approach, LLM Agent-based Asset Pricing Models (AAPM), which fuses qualitative discretionary investment analysis from LLM agents and quantitative manual financial economic factors to predict excess asset returns. The experimental results show that our approach outperforms machine learning-based asset pricing baselines in portfolio optimization and asset pricing errors. Specifically, the Sharpe ratio and average $|\alpha|$ for anomaly portfolios improved significantly by 9.6% and 10.8% respectively. In addition, we conducted extensive ablation studies on our model and analysis of the data to reveal further insights into the proposed method.

1 Introduction

The pricing of financial assets, such as stocks, has been a focal point in empirical financial economics research. It has a significant impact on social good by moving towards Pareto efficiency in capital allocation. Current asset pricing methods rely on carefully crafting manual macroeconomic indicators or company-specific factors as predictors of future excess returns (Fama and French, 1992, 2015). Despite its great success in the real-world market, they have been challenged by the Efficient Market Hypothesis (EMH) that manual factors will ultimately lose their predictive power in an efficient market when these predictors are fully discovered and used by market participants.

Due to this rationale, linguistic data, which are the primary sources of traditional discretionary investing, become essential. This is because the dynamics of society and the market are largely driven by the information flow of language. This is also evident in the real financial world, where discretionary portfolio management remains significant today (Abis, 2020). Such investment decisions are mainly shaped by the manager's experience and intuition, as they evaluate assets and determine their value based on information from news, investigations, reports, etc., instead of depending on quantitative models.

This phenomenon highlights two key points. First, qualitative discretionary analysis can uncover valuable pricing insights that are absent in economic indicators or market data. Second, even with the integration of current NLP and semantic analysis methods, quantitative factor models have not fully captured these insights. Achieving the synergy between both remains a complex yet appealing objective (Cao et al., 2021). Nonetheless, leveraging linguistic information is complicated as it requires financial reasoning and long-term memory of tracking events and company impressions to interpret. Furthermore, suboptimal interactions in model design between linguistic and manual factors can end up as noise (Bybee et al., 2023).

In this study, we introduce a novel asset pricing approach, LLM Agent-based Asset Pricing Models (AAPM), which fuses discretionary investment analysis simulated by an LLM agent and quantitative factor-based methods. AAPM employs the LLM agent to iteratively analyze the latest news, supported by a memory of previous analysis reports and a knowledge base comprising books, encyclopedias, and journals. The embedding of analysis reports is merged with manual factors to predict future excess asset returns. Besides offering a performance edge, our method also provides enhanced interpretability through generated analysis reports. We evaluate our approach with a dataset consisting of two years of news and approximately 70 years of economic and market data. The experimental results show that our approach surpasses machine learning-based asset pricing baselines, achieving a 9.6% increase in the Sharpe ratio and a 10.8% improvement in the average $|\alpha|$ for asset pricing errors in character-section portfolios. Our primary contributions are summarized as follows:

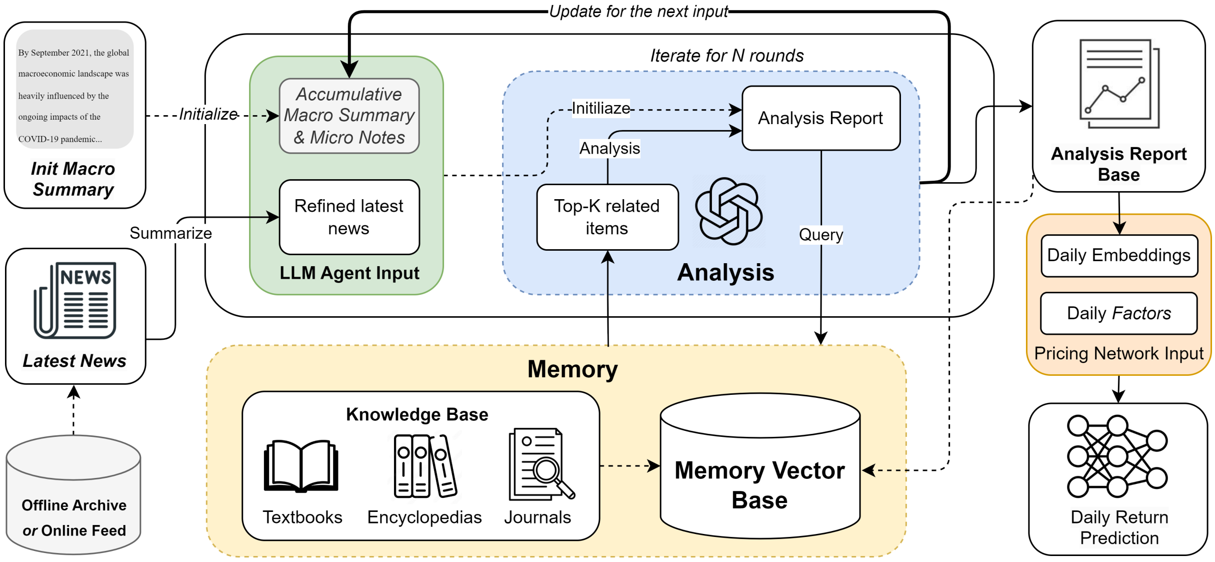

Figure 1: The LLM agent produces analysis report from the latest news through a multi-step refinement, incorporating past reports and domain knowledge from memory. For simplicity, the filter for irrelevant news is excluded. A macro and micro note, continuously updated by the latest analysis report, is used to provide additional context. The average embedding of daily analysis reports will be input into the pricing network along with daily manual factors.

- Introduced a novel LLM agent architecture to analyze business news for discretionary investment insights as pricing signals.
- Proposed a hybrid asset pricing framework combines qualitative discretionary analysis and quantitative manual factors.
- Performed comprehensive experiments to assess the effectiveness of the proposed approach with in-depth analysis of components.

Our code and data are provided in the supplementary material and will be made public after the double-blind review process.

2 Related Work

2.1 Asset Pricing for Security

Asset pricing aims to search for the fair price of financial assets, such as securities. Sharpe (1964) introduced the groundbreaking Capital Asset Pricing Model (CAPM), which breaks down the expected return of an asset into a linear function of the market return. Various extensions of the CAPM have been developed. Merton (1973) incorporated wealth as a state variable, while Lucas Jr (1978) considered consumption risk as a pricing factor. The single-factor CAPM was later expanded into multi-factor models. Fama and French (1992) proposed the Fama-French 3-factor (FF3) model, which explains returns by size, leverage, book-to-market equity, and earnings-price ratios. They later revised it to a 5-factor model (Fama and French, 2015). Furthermore, Carhart (1997) identified momentum as an additional factor. Ross (1976) formulated the Arbitrage Pricing Theory (APT), which considers asset pricing as an equilibrium in the absence of arbitrage opportunities. The Stochastic Discount Factor (SDF) calculates the price by discounting future cash flows using a stochastic pricing kernel (Cochrane, 2009).

2.2 Financial Machine Learning

The application of machine learning techniques has been introduced to explore the non-linear interactions among the growing "factor zoo" (Feng et al., 2020). Instrumented Principal Component Analysis (IPCA) was developed by Kelly et al. (2020) to estimate latent factors and their loadings from data. Gu et al. (2020) introduced a deep neural network to model interactions. Gu et al. (2021) proposed a conditional autoencoder that considers latent factors and asset characteristics as covariates. Chen et al. (2024) utilized Generative Adversarial Networks to train a neural SDF based on the methods of moments. Additionally, Bybee et al. (2021) conducted an analysis of the Wall Street Journals (WSJ) to gauge the state of the economy. Based on this analysis, Bybee et al. (2023) further suggested using Latent Dirichlet Allocation

(LDA) to analyze monthly news topics from WSJ as pricing factors. Recent NLP methods (Xu and Cohen, 2018; Xie et al., 2022) have been employed to forecast stock movements, in contrast to asset pricing, they do not aim to find interpretable factors that explain anomalies in excess asset returns. Our LLM-based approach offers an alternative interpretation through analysis reports.

2.3 Large Language Model Agents

LLM agents possess powerful emergent capabilities, such as reasoning, planning, and tool-using (Achiam et al., 2023). The core of LLM agent programming lies in prompting, which employs contextual hinting text to regulate the output of LLM (Liu et al., 2023). Several prompting strategies have been proposed. Chain-of-Thoughts (CoT) (Wei et al., 2022) encourages the agent to reason in a step-by-step manner. Yao et al. (2022) introduced the ReAct prompt, enabling the agent to refine its output based on the results of previous attempts. It allows the agent to use external tools, such as databases and search engines. Memory is another crucial component of LLM agents. Hu et al. (2023) introduced databases as symbolic memories. Packer et al. (2023) stores dialogues in both long- and short-term memory, analogously to operating systems. Cheng and Chin (2024) developed an agent capable of making "investment" decisions on social science time series based on input news, reports, etc., and knowledge base, as well as the Internet. We focus on using the agent to simulate discretionary investment decision-making to synergize qualitative and quantitative asset pricing.

3 Method

Given a state vector $V_{\tau,a}$ at a time point $\tau \in \{0, 1, 2, ...\}$, which represents the current status of the market, society, and an asset a, an asset pricing model predicts the excess returns $r_{\tau+1,a}$ of the asset at the subsequent time point, expressed as $P(r_{\tau+1,a}|V_{\tau,a})$. In our study, each time point corresponds to one day. In traditional factor-based methods, the state $V_{\tau,a} \in \mathcal{N_F}^N$ is a vector composed of N_F factors that are manually derived from economic indicators, market data, asset characteristics, etc. For instance, the market excess return, the performance disparity between small and large firms, and the difference between high and low book-to-market companies in the Fama-French 3-factor model (Fama and French, 1992). Recently, Bybee et al. (2021) demonstrated that a collection of business news can serve as an alternative representation of macroeconomic conditions, while Bybee et al. (2023) employs LDA to extract news characteristics as economic predictors for pricing. Building on this idea, we use the average embedding of analysis reports that mine values from the news as a proxy for the society, economic, and market states.

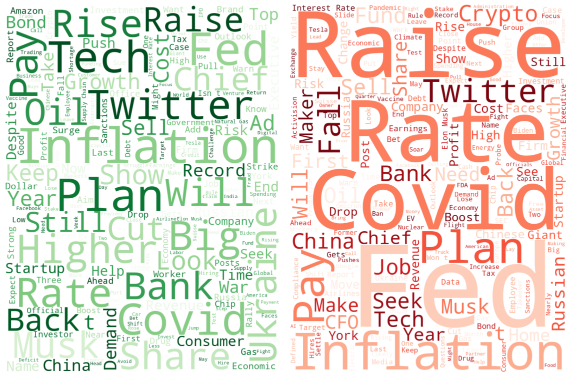

Figure 2: Visualization of the key words in the titles of news articles on the days when the market return is positive (left) and when it is negative (right).

Business news in major media outlets like the WSJ carries important market information, however they typically restrict their interpretations and opinions, leaving room for discretionary analysis. It is crucial to understand that business events are often interrelated.

As visualized in Figure 2 about the keywords found in the titles of the news articles on days with positive and negative market returns. It corresponds well with human intuition about how the market trend was driven, long-term events like the FED rate hike, COVID, and inflation worries have had the most significant negative effects on the market over the two-year span from Sep. 2021 to Sep. 2023 of our dataset, whereas elements such as technology, Twitter, and inflation control measures have driven market growth. Interpreting business news about such key events requires an extrapolation process that depends on extensive background knowledge and historical events.

Based on these observations, we introduce AAPM, utilizing an LLM agent with long-term memory of domain knowledge and historical news analysis to iteratively analyze the input news and generate the analysis report, as detailed in Section 3.1. Subsequently, we combine these qualitative analysis reports and quantitative manual factors

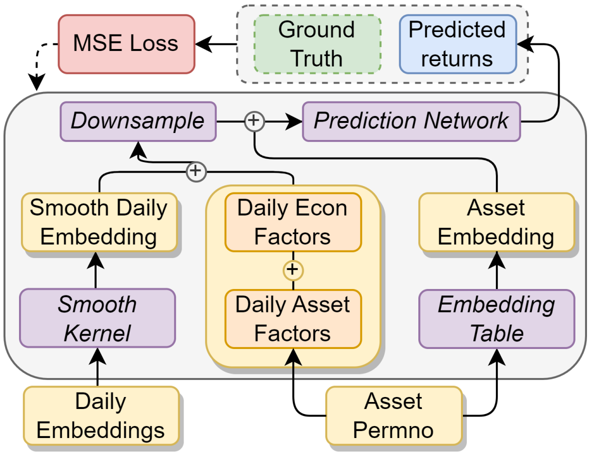

Figure 3: The demonstration of our hybrid asset pricing network. The purple boxes mark the computational components. Yellow boxes are data, the circled plus symbol means contatenation. The MSE loss computed with predicted returns feedback to update the network.

to feed into our hybrid asset pricing network in Section 3.2.

3.1 Discretionary Analysis with LLM Agent

The agent utilizes the latest news x_t at time t (e.g., a WSJ article published at 9:32 AM on 6 June 2020), along with a note n_t on macroeconomics and market trends, to generate an analysis report R_t. The note n_t is initalized with a macroeconomics summary n_0 produced by GPT-3.5-Turbo-1106 (Brown et al., 2020), the LLM used in our study, prior to its knowledge cut-off date d_k. It offers necessary macroscopic context on economic and societal trends not directly available from the news or the memory. The note is then iteratively updated to $n_{t'}$ with the new analysis report R_t to keep the context up-to-date, and we also prompt the agent to document investment ideas and market thoughts in the notes to provide a short-term background such as the trends on the market, long-term research oppurtunities to watch. To ensure the note is continuously updated without missing information while preventing information leakage, the dataset in our study starts from d_k+1, immediately following the knowledge cut-off date.

The analysis process begins with generating a refined news item x'_t that summarizes key information from the raw input x_t. This step helps control the input length and standardizes the format and style. The refined news x'_t and the note n_t are then combined to form an input I_t for the agent. The agent will determine if the news contains investment information: if not, it will be skipped; otherwise, an initial analysis report R^0_t will be created. The report undergoes iterative refinement over N rounds. In each round i, the report R^{i-1}_t is used to query an external memory M^t, a vector database initialized with the SocioDojo knowledge base (Cheng and Chin, 2024), which includes textbooks, encyclopedias, and academic journals in fields such as economics, finance, business, politics, and sociology. We use BGE (Xiao et al., 2023) as the embedding model f_e, which maps text to a vector $e \in \mathcal{R}^{d_{emb}}$ for querying the memory. This choice is based on the MTEB leaderboard (Muennighoff et al., 2022), where we selected the best retrieval model considering performance, model size, and embedding vector length. In each round i, the top-K most relevant items $\{m^{t,i}_j\}^K_{j=1} \subset M^t$ are retrieved and provided to the agent along with the report R^{i-1}_t to produce the refined report R^i_t. The report R^N_t generated after the N-th round is used as the final analysis report R_t for the news x_t and to update the note as $n_{t'}$. Then it is inserted into the memory M^t for future reference and pricing, updating the memory to $M^{t'}$.

The pricing network will utilize the analysis reports $\{R_{t^d_i}\}^{N_d}_{i=1}$ of all filtered news $\{x_{t^d_i}\}^{N_d}_{i=1}$ for a given day d, where N_d represents the number of filtered news items on day d. Figure 1 provides an overview of the entire analysis process. The prompts employed in our agent are detailed in Appendix E. In Section 4.4, we conduct experiments on our agent design and the impact of N and K.

3.2 Hybrid Asset Pricing Network

We use the embedding model f_e to transform each report $R_{t^d_i}$ into an embedding $e_{t^d_i}$, where t^d_i represents the timestamp of the i-th news on day d. The average embedding of the analysis reports on a given day d is calculated as $e_d = \sum^{N_d}_{i=1} e_{t^d_i}/N_d$. According to Bybee et al. (2023), a single day's news is insufficient to fully capture the broader economic and market conditions. Therefore, we employ a sliding window of L_W to derive a ***smoothed daily embedding*** s_d using the average embeddings of the most recent $L = min(L_W, d)$ days $\{e_{d-L+1}, e_{d-L+2}, ..., e_d\}$ as follows:

$$s_d = \sum_{i=1}^{L} \kappa(L,i) e_{d-L+i}$$

where $\kappa(L,i)$ is an exponential decay kernel defined as $\frac{\eta^{L-i}}{\sum^L_{j=1}\eta^{L-j}}$. The decay coefficient is denoted as $0 < \eta < 1$. We form a raw hybrid state $h_{d,a} = [s_d; v_{d,a}]$ by concatenating the smoothed daily state s_d with a vector $v_{d,a} \in \mathcal{R}^{N_F}$ of N_F manual-constructed financial economic factors. The asset a is indexed by a permanent number (permno) from the Center for Research in Security Prices (CRSP) [1] database. The hybrid state is subsequently downsampled by $h'_{d,a} = \sigma(W_S h_{d,a})$, where σ denotes the ReLU function and $W_S \in \mathcal{R}^{d_{model} \times (d_{emb} + N_F)}$ is a parameter matrix.

To capture the asset-specific loading to the hybrid state especially to the asset-agnostic s_d, we define an asset embedding $E \in \mathcal{R}^{N_A \times d_{model}}$, which can be looked up via the permnos of the assets. Here, N_A denotes the total number of assets and d_{model} is the dimension of the embedding. We then concatenate the asset embedding E_a with the downsampled hybrid state to form $\hat{h}_{d,a} = [h'_{d,a}; \sigma(E_a)]$, the **asset-specfic hybrid state** for a.

The excess return of asset a for the next day is predicted by $r_{d+1,a} = f_P(\hat{h}_{d,a})$, where $f_P = f_{P_{inp}} \circ f_{H_1} ... \circ f_{P_{out}}$ represents a multi-layer fully connected prediction network. Specifically, $f_{P_{inp}}(\cdot) = \sigma(W_{P_{inp}} \cdot)$, with $W_{P_{inp}} \in \mathcal{R}^{2d_{model} \times d_{model}}$, and $f_{P_{out}}(\cdot) = W_{P_{out}} \cdot$, where $W_{P_{out}} \in \mathcal{R}^{d_{model} \times 1}$. Additionally, f_{H_k}, for $k \in [1, 2, 3, ...]$, denotes hidden layers parameterized by $W_{H_k} \in \mathcal{R}^{d_{model} \times d_{model}}$. For simplicity, batch normalizations, residual connections, and dropout layers are not included. Figure 3 illustrates the prediction network.

The hybrid asset pricing network, represented as f_H and parameterized by θ, comprises the embedding table E, the downsampling matrix W_S, and the prediction network f_P. We train f_H using the Mean Square Error (MSE) criterion, which minimizes the average squared difference between the predicted return $r_{d+1,i}$ and the ground truth $\hat{r}_{d+1,a}$ over the training set, written as

$$\arg\min_{\theta} \frac{1}{N_A N_D} \sum_{d,a} (r_{d+1,a} - \hat{r}_{d+1,a})^2, \quad where \; r_{d+1,a} = f_H(h_{d,a}; \theta)$$

Where N_D denotes the number of days in the training set. The model is trained for T episodes with a batch size of B. We initially pre-train this hybrid asset pricing network f_H to make use of the historical factor data available before the beginning of the news dataset. During this pre-training phase, a placeholder embedding (such as the embedding for the word "Null") is utilized.

[1] https://www.crsp.org/

4 Experiment

We conduct experiments to assess the asset pricing efficacy of the proposed AAPM. The experimental setup is detailed in Section 4.1. Subsequently, we present the outcomes of the portfolio optimization experiments in Section 4.2 and the asset pricing error in Section 4.3. An extensive ablation study of our method is provided in Section 4.4. Furthermore, we explore the predictive capabilities of refined news on economic indicators and stock movements in Appendix C.

4.1 Experiment Setting

We build a dataset comprising two years of WSJ articles spanning from September 29, 2021, to September 29, 2023, following the knowledge cut-off of the version of GPT we used. This approach mitigates potential information leaks while maintaining continuity in note n. Besides the LLM filtering described in Section 3.1, we also manually excluded articles on unrelated topics like travel, lifestyle, and puzzles, based on their WSJ categories. Visualizations of our news dataset can be found in Appendix B. The daily asset returns are sourced from CRSP, while daily risk-free returns and market returns are obtained from Kenneth French's data library [2].

We construct financial economic factors following Jensen et al. (2023). In line with Chen et al. (2024), we duplicate the values from the previous time step for factors that are not updated in the current step to handle discrepancies in the update frequencies of the factors. Additionally, we imputed the missing data values using the cross-sectional median. The data split remained consistent across all our experiments: the initial 9 months of data were utilized as the training set, the following 3 months served as the validation set, and the last 1 year was reserved for testing.

We select five recent asset pricing baselines from highly reputed financial economics journals, validated under current empirical finance standards, as indicated by Jensen et al. (2023), to assess our approach: NN (Gu et al., 2020) introduced a deep neural network for asset pricing; IPCA (Kelly et al., 2020) developed an instrumental PCA to identify hidden factors and loadings; CA (Gu et al., 2021) proposed to use a conditional autoencoder; NF (Bybee et al., 2023) employs LDA for the WSJ news as hidden factors similar to ours; and CPZ

[2] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french

	SR ↑			MDD (%) ↓		
	TP	EW	VW	TP	EW	VW
NN	3.82	2.83	2.36	4.82	8.12	9.12
IPCA	4.07	2.96	2.66	3.77	5.77	8.63
CA	4.03	2.85	2.55	3.79	6.31	**4.66**
NF	*3.73*	*2.76*	*2.34*	*5.12*	*7.91*	*6.31*
CPZ	4.10	3.02	2.61	4.32	6.27	5.71
Ours	4.38	3.29	3.01	**3.66**	5.64	5.17
w/ G.4	**4.45**	**3.43**	**3.09**	3.82	**5.57**	4.77

Table 1: Sharpe Ratio (SR) and Maximal Drawdown (MDD) for Tengency Portfolio (TP), Equal-Weighted (EW) and Value-Weighted (VW) long-short portfolio built based on NN (Gu et al., 2020), IPCA (Kelly et al., 2020), CA (Gu et al., 2021), NF (Bybee et al., 2023), CPZ (Chen et al., 2024), and our method with the default GPT-3.5 or GPT-4. We bolded the best results and underlined the second bests.

(Chen et al., 2024) utilized GAN to address stochastic discount factors. We replicated these models using the configurations from their respective papers with their carefully chosen factor sets. For both our method and the baselines, we performed a hyper-parameter search to compare the best results. The hyper-parameter optimization setting for our method is detailed in Appendix A.

4.2 Portfolio Optimization

We begin by testing the Sharpe ratio for portfolios built on the predicted returns of individual assets. The Sharpe ratio (SR) (Sharpe, 1998) quantifies the risk-adjusted performance of a portfolio as $S_p = \frac{\bar{r}_p - \bar{r}_f}{\sigma(r_p)}$, where r_f stands for the risk-free return, r_p represents the portfolio return and σ indicates the standard deviation. Furthermore, we evaluate the maximum drawdown, which is the largest decrease in the total value of the portfolio up to time T, expressed as $MDD(T) = \max_{\tau\in(0,T)}[\max_{t\in(0,\tau)} X(t) - X(\tau)]$. Here, $X(\tau)$ is the highest value and $X(t)$ is the lowest value of the portfolio within the time interval $(0, \tau)$.

We evaluate three prevalent methods for portfolio construction. The Tangency Portfolio (TP), where the asset weights are calculated as $w_t = E_t[R^e_{t+1} {R^e_{t+1}}^T]^{-1} E_t[R^e_{t+1}]$, with R^e_{t+1} denoting the predicted excess returns of all assets. Provides a theoretical portfolio in an ideal market without trading frictions. Next, we examine the more practical long-short decile portfolios, which involve ranking assets by their expected returns, going long on the top decile, and shorting the bottom decile. The assets in these portfolios can be either "Equally-Weighted" (EW) or weighted by their market capitalization, known as "Value-Weighted" (VW).

The experiment results are presented in Table 1. Our approach achieved the highest SR across all three portfolios, with SR improvements of 6.8%, 8.9%, and 13.2% respectively over the best baseline methods (CPZ for TP and EW, IPCA for VW), averaging a 9.6% increase. Additionally, it secured the best or second-best MDD in TP and EW compared to the leading baseline IPCA, with gains of 2.9% and 2.3% respectively. In VW, the MDD underperforms the top baseline CA by 10.9%. However, substituting GPT-3.5 in our model with GPT-4-0613 which has the same knowledge cutoff, resulted in SR improvements of 8.5%, 13.6%, and 16.2% across the three portfolios, and improved MDD levels to gains of 1.3%, 3.5%, and -2.4% relative to the best baselines.

	avg $\|\alpha\|$	avg $\|t(\alpha)\|$	$\frac{\#\|t(\alpha)\|>1.96}{\#test\ assets}$	GRS
NN	0.83	2.89	0.64	6.89
IPCA	0.76	2.45	0.55	6.38
CA	0.77	2.63	0.52	6.42
NF	*0.89*	*2.77*	*0.62*	*7.32*
CPZ	0.74	2.44	0.49	6.77
Ours	0.66	2.40	**0.46**	6.34
w/ G.4	**0.64**	**2.36**	**0.46**	**6.28**

Table 2: Asset pricing errors for anomaly portfolios with NN (Gu et al., 2020), IPCA (Kelly et al., 2020), CA (Gu et al., 2021), NF (Bybee et al., 2023), CPZ (Chen et al., 2024), and our method with GPT-3.5 and GPT-4. We bolded the best results and underlined the second bests.

4.3 Asset Pricing Error

We further analyze the asset pricing errors of the proposed method. Following Bybee et al. (2023), we chose 78 anomaly portfolios as test assets. These portfolios were constructed using 78 characteristics, including typical anomaly characteristics such as idiosyncratic volatility, accruals, short-term reversal, and others, as identified by Gu et al. (2020). We applied multiple metrics. The average absolute alpha $avg.|\alpha|$ is computed by dividing the expected value of the estimated error term $\hat{\epsilon_{t,i}}$ by the square root of the average squared returns $E[R_{t,i}]$ for all quantile-sorted portfolios. This normalization was performed to account for variations in average returns between portfolios. To measure statistical significance, we calculated the average t-value for the results and analyzed the proportion of t-values exceeding 1.96. Moreover, we conducted a Gibbons, Ross, and Shanken (GRS) test (Gibbons et al., 1989) to determine if the regression intercepts, represented by $\alpha_1, \alpha_2, ..., \alpha_n$, are collectively zero. This test helps to evaluate the overall significance of the intercepts in the regression analysis.

Table 2 displays the results. Our method secured either the top or second-best performance among all benchmarks. It demonstrates a 10.8% and 13.5% reduction in average $|\alpha|$ for GPT-3.5 and GPT-4 respectively when compared to CPZ, the leading benchmark, along with a 1.6% and 3.3% increase in t-value. Additionally, there is a 6.1% reduction in the proportion of pricing results with a t-value exceeding 1.96 compared to CPZ for both GPT-3.5 and GPT-4, as well as a 0.6% and 1.6% enhancement in the GRS test compared to IPCA.

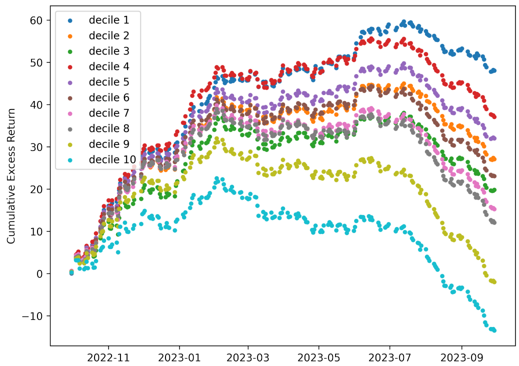

Figure 4: Cumulative excess return for decile portfolios.

We move forward to assess the proposed method by applying it to the pricing of decile portfolios. This process includes sorting the assets according to their predicted returns and then forming portfolios for each decile. Figure 4 shows the cumulative excess return over time. The figure clearly demonstrates that each decile creates a distinct ranking of returns in the right position, suggesting that the proposed approach effectively predicts returns at various levels.

4.4 Ablation Study

We conduct ablation studies to examine the influence of various components in our approach. Initially, we evaluate the performance of different modules in our agent design in Section 4.4.2, followed by an examination of the depth and width of the analysis, which are controlled by N and K respectively, in Section 4.4.2.

	SR	MDD	avg $\|\alpha\|$	avg $\|t(\alpha)\|$
NF	*2.76*	*7.91*	*0.89*	*2.77*
+ Factors	*2.66*	*8.82*	*0.97*	*2.86*
Naive	2.82	6.03	0.88	2.72
+ RAG	2.94	5.89	0.83	2.66
+ Emb.	2.88	6.42	0.86	2.71
Memory	2.99	6.99	0.81	2.64
+ Factors	3.03	7.12	0.79	2.62
Hybrid	3.14	5.59	0.73	2.49
+ Refine	3.26	6.31	0.70	2.46
+ Notes	3.18	6.91	0.74	2.55
Ours	**3.29**	**5.64**	**0.66**	**2.40**

Table 3: Ablation study of AAPM and comparison with NF (Bybee et al., 2023). "Naive" directly produce the analysis report given news and only daily embeddings are inputted to the pricing network. "+ RAG" introduces the external memory and retrieves Top-K items when performing analysis. "+ Emb." introduces the asset embeddings. "Memory" baseline incorporate both "+ RAG" and "+ Emb." "+ Factors" introduces the daily manual factors into the pricing network in "Memory". "Hybrid" baseline pretrained the pricing network of "Memory". "+ Refine" refines the analysis report iteratively in N rounds for "Hybrid". "+ Notes" introduces the macro economics and micro market notes. "Ours" is our method that combines "+ Refine" and "+ Notes" in "Hybrid".

4.4.1 Agent Architecture Design

We analyze our architecture in a reverse manner, beginning with a "Naive" agent that generates the analysis report directly from the refined news without any supplementary information or iterative analysis, while the pricing network solely uses the daily embeddings as input. We then incrementally add components to develop stronger baselines until arriving at our method. The results are shown in Table 3, and the baseline illustrations are provided in Appendix D.

Furthermore, we contrast these methods with the news-based asset pricing baseline NF (Bybee et al., 2023), along with an NF model incorporating manual factors, akin to our full model. It is important to highlight that NF employed WSJ news over a period of 33 years, whereas we utilized only 2 years of news data.

Owing to the analytical capabilities and feature

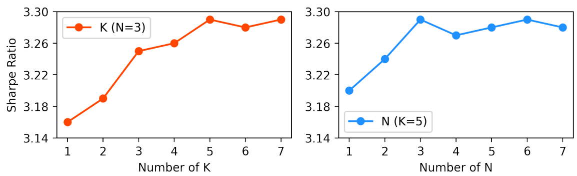

Figure 5: The Sharpe ratio of equal-weighted portfolios given different number of K and N.

extraction proficiency of LLMs, the "Naive" baseline enhances the SR by 2.2% with comparable pricing errors to NF. Incorporating external memory further boosts the SR by 4.3% and decreases the average $|\alpha|$ by 5.7% over "Naive", highlighting the significance of additional contextual information when interpreting business news. Moreover, asset embedding contributes to a 2.1% increase in SR and a 2.3% reduction in average $|\alpha|$ by introducing asset-specific loadings.

By combining both, the "Memory" baseline enhances the SR of "Naive" by 6.0% and decreases the average $|\alpha|$ by 8.0% with a lower t value. Incorporating the manual factors, the SR saw a slight increase of 1.3%, while the average $|\alpha|$ decreased by 2.5%. In comparison, the performance of NF declined after the introduction of manual factors, which is consistent with the findings of Bybee et al. (2023), where the inclusion of Fama-French factors reduces the SR, which may due to suboptimal interactions between factors and news features.

After pretraining the pricing network with historical factor data, the performance of the "Hybrid" baseline saw a notable enhancement of 5.0% in SR and a 9.9% reduction in the average $|\alpha|$ when compared to the "Memory" baseline. This demonstrates the synergy between manual factors and LLM-generated reports, resulting in a successful non-linear interaction. The improvements from our iterative refinement and long-term notes over the "Hybrid" baseline are 3.8% and 1.3% in SR, 4.1% and a slight negative -1.4% in average $|\alpha|$, respectively with a lower t value and a similar level of MDD. These enhancements collectively yield 4.8% and 9.6% gains in SR and average $|\alpha|$ respectively in our full method compared to a "Hybrid" baseline, underscoring the effectiveness of our agent architecture design.

4.4.2 Analysis Depth and Width

We further investigate the depth of the analysis, which is controlled by the number of iterations N to refine the analysis report, and the width, which is determined by K, the amount of relevant information to check. The results are shown in Figure 5. We keep one variable constant and test the other. We observe that the agent benefits from more rounds of analysis and a broader range of relevant information overall with a sharp decline in marginal gain after a certain point around $K \times N = 15$, likely due to the sufficiency of the provided information. Thus, we test an extreme case where $N = 1$ and $K = 15$, resulting in the SR dropping to 3.12. This indicates that iterative refinement is necessary, as items retrieved in different rounds of refinement provide diverse information as the query evolves. In contrast, a single retrieval leads to items falling into the same topic, with the value of additional items decreasing rapidly and potentially introducing noise.

5 Discussion

Our proposed approach presents a promising method to fuse qualitative discretionary investment with quantitative factor-based strategies through the use of LLM agents. Nonetheless, there is still much to investigate regarding additional capabilities of LLM agents that could further enhance asset pricing power. Firstly, internet access and a broader range of information sources, including those available in SocioDojo, may enable the agent to generate more in-depth analyses, as discretionary investment relies on information beyond just news or domain knowledge. Secondly, employing specialized financial LLMs like FinGPT (Yang et al., 2023) could further improve the agent's financial analytical capabilities. Finally, it is crucial to consider multimodal information, such as diagrams and figures, which are frequently presented in financial documents.

6 Conclusion

In this research, we introduced AAPM, a model that combines qualitative analysis from the LLM agent with quantitative factors in asset pricing. AAPM surpassed established asset pricing methods in multiple evaluations, including portfolio optimization and asset pricing error. Additionally, we performed an in-depth analysis of each component in our agent design. We believe that our study can improve the comprehension of the interaction between discretionary investment and quantitative factor-based models, toward a society with increased economic efficiency.

Limitations

Our experiments only focus on the US market and English news, which may potentially impact model performance in lower-resources languages. In order to exclude the information leak, we can only apply news data after September 2021 which restricts our study to a 2 years period after this time, however, we use a large test split where half of the dataset was applied as the test set to best evaluate how well the proposed method can be generalized beyond the training period. Finally, public information in the stock market includes not only news, but also reports, reports from social networks, academic journals, opinions from experts, etc.; we do not cover these information nor consider multimodal inputs as discussed in Section 5.

Ethics Statement

We do not identify any ethical concerns in our approach. Our study does not involve any human participation. Furthermore, the application area of our method is not directly related to humans, reducing the risk of abuse or misuse. In fact, considering a wider range of information, our method has the potential to enhance market efficiency, resulting in economic benefits for society.

References

Simona Abis. 2020. Man vs. machine: Quantitative and discretionary equity management. *Machine: Quantitative and Discretionary Equity Management (October 23, 2020)*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Leland Bybee, Bryan Kelly, and Yinan Su. 2023. Narrative asset pricing: Interpretable systematic risk factors from news text. *The Review of Financial Studies*, 36(12):4759–4787.

Leland Bybee, Bryan T Kelly, Asaf Manela, and Dacheng Xiu. 2021. Business news and business cycles. Technical report, National Bureau of Economic Research.

Sean Cao, Wei Jiang, Junbo L Wang, and Baozhong Yang. 2021. From man vs. machine to man+ machine: The art and ai of stock analyses. *Columbia Business School Research Paper*.

Mark M Carhart. 1997. On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82.

Luyang Chen, Markus Pelger, and Jason Zhu. 2024. Deep learning in asset pricing. *Management Science*, 70(2):714–750.

Junyan Cheng and Peter Chin. 2024. Sociodojo: Building lifelong analytical agents with real-world text and time series. In *The Twelfth International Conference on Learning Representations*.

John Cochrane. 2009. *Asset pricing: Revised edition*. Princeton university press.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Eugene F Fama and Kenneth R French. 1992. The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465.

Eugene F Fama and Kenneth R French. 2015. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. 2020. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370.

Michael R Gibbons, Stephen A Ross, and Jay Shanken. 1989. A test of the efficiency of a given portfolio. *Econometrica: Journal of the Econometric Society*, pages 1121–1152.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2021. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450.

Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. 2023. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*.

Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. 2023. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518.

Bryan Kelly, Dacheng Xiu, et al. 2023. Financial machine learning. *Foundations and Trends® in Finance*, 13(3-4):205–363.

Bryan T Kelly, Seth Pruitt, and Yinan Su. 2020. Instrumented principal component analysis. *Available at SSRN 2983919*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Robert E Lucas Jr. 1978. Asset prices in an exchange economy. *Econometrica: journal of the Econometric Society*, pages 1429–1445.

Robert C Merton. 1973. An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, pages 867–887.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*.

Stephen A Ross. 1976. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360.

William F Sharpe. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.

William F Sharpe. 1998. The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management*, 3:169–185.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Yong Xie, Dakuo Wang, Pin-Yu Chen, Jinjun Xiong, Sijia Liu, and Sanmi Koyejo. 2022. A word is worth a thousand dollars: Adversarial attack on tweets fools stock predictions. *arXiv preprint arXiv:2205.01094*.

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *FinLLM Symposium at IJCAI 2023*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

A Hyperparam Search

Parameter	Distribution
Learning rate	{1e-3,1e-4,5e-4,5e-3}
d_{model}	{128,256,512,768,1024}
d_{emb}	{128,256,512,768,1024}
Epochs	{50,100,150,200}
Hidden Layers	{0,1,2,3,4,5}
Dropout rate	$U(0, 0.3)$
Batch size	$U_{log}(32, 1024, 8)$
η	$U(0.9, 1)$
L_W	{1,7,15,30,45,60,90,180}
N	{1,2,3,4,5}
K	{1,2,3,4,5}

Table 4: Distributions for the key hyperparameters in the hyperparameter search.

For our approach, we conduct hyperparameter searches using Weights & Biases Sweep (Biewald, 2020). Table 4 shows the distribution of empirically significant parameters used for our hyperparameter search. Here, $U(a, b)$ signifies a uniform distribution between a and b, while $U_{log}(a, b, r)$ indicates a logarithmic uniform distribution with base r between a and b. The evaluation criteria of our method are based on the Sharpe ratio of an equal-weight long-short portfolio.

We conducted our experiments on our clusters, the major workload has the following configuration:

- 2 × Intel Xeon Silver 4410Y Processor with 12-Core 2.0GHz 30 MB Cache
- 512GB 4800MHz DDR5 RAM
- 2 × NVIDIA L40 Ada GPUs (no NVLink)

We employed PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019) for parallel training.

B Dataset visualizations

Figure 6 illustrates the variations in the number of articles and assets over time. We analyze the primary topics discussed in the news articles within our dataset across different periods. The topics were determined based on the titles of the news articles for each season. Common words such as "US," "Stock," and "Market" were excluded as they did not effectively represent the event's topic. The

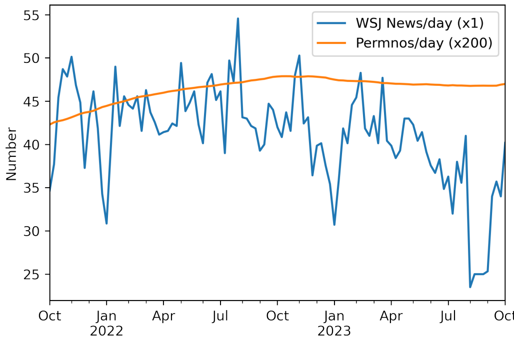

Figure 6: The number of filtered WSJ articles and active assets per day.

resulting word cloud is shown in Figure 7. It is clear that the economy is mainly influenced by various long-term events. It begins with a gradual decline in the emphasis on COVID. Then, the focus shifted towards managing inflation and the decisions made by the FED. The banking crisis at the start of 2023 soon became the new central point, followed by the acknowledgment of AI as a key driver for the economy, mainly due to the success of LLMs. This indicates that these event trends have the potential to serve as strong predictors of economic indicators and the market. This is also reflected in Appendix C, where we evaluated that news articles have significant predictive power for economic indicators and market trends.

We then use GPT to analyze the relevant tickers for each news item in our dataset with the following prompt:

> You are a helpful assistant designed to analyze the business news. You need to extract the stock tickers of the companies most closely related to the news. If there is no relevant ticker, return an empty list. You should never make up a ticker that does not exist. Now, analyze the following news: {input}

The stock tickers linked to the news in our dataset are displayed in Figure 8. Over the two-year span, technology stocks have evidently been the market's primary focus, aligning with our impression and the actual robust performance of these stocks over the period.

C News as Financial Economic Predictor

To explore the predictive capability of business news on financial and economic dynamics, we conduct an experiment using refined news features to forecast the economic indicators in Appendix C.1 and market movements in Appendix C.2. We embed the refined news directly and use the daily averaged embeddings of the refined news as predictors in our experiments.

C.1 Economic Indicators

We assess the capability of news features to forecast the daily percentage changes in typical and most popular macroeconomic indicators in different topics sourced from the FRED database[3]. These indicators encompass the stock market (SP500), the market yield on U.S. Treasury Securities at a 10-Year constant maturity (DGS10), Moody's seasoned Baa corporate bond minus the federal funds rate (BAAFF), the 10-year breakeven inflation rate (T10YIE), Brent crude oil prices (DCOILBRENTEU), and the 30-year fixed-rate conforming mortgage index (OBMMIC30YF). The findings are illustrated in Figure 9. The forecasted results exhibit a high degree of accuracy, as evidenced by the high R2 score. This implies that news provides valuable insights for predicting macroeconomic indicators.

C.2 Stock Price Predictor

We further investigate the predictive power of news features to the price movements of individual stocks. We chose 8 typical stocks that has been frequently mentioned in the new from our analysis in Appendix 8, and used refined news features as predictors to estimate their daily percentage price changes. The results are displayed in Figure 10, with the corresponding R2 scores given in brackets. We note high accuracy and R2 scores for all selected stocks, suggesting that news can significantly help in forecasting stock prices. However, it is important to recognize that due to non-stationarity and the risk of overfitting, stock price prediction cannot be directly applicable as asset pricing (Kelly et al., 2023), but it provides insights into the value of the news in the pricing of individual stock.

D Illustration of the Ablation Baselines

We progressively developed three baselines, starting with a naive agent, followed by a memory agent enhanced with an external vector base, and culminating in a hybrid agent that incorporated manual

[3] https://fred.stlouisfed.org/

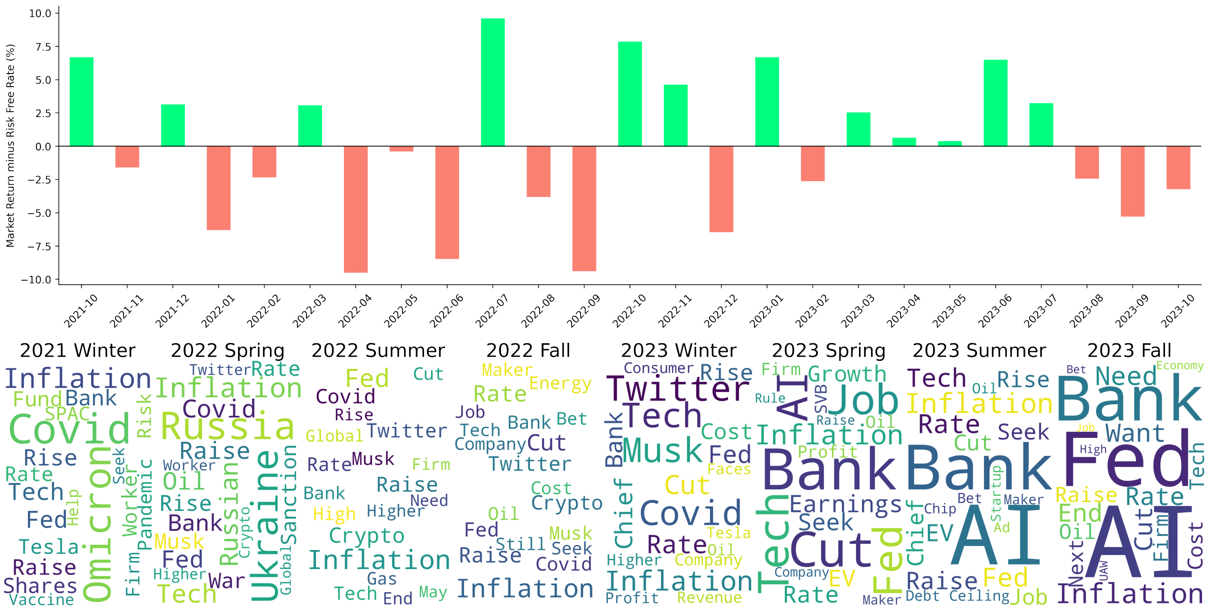

Figure 7: The word cloud of topics of the WSJ business news over time in our dataset (Bottom) compared to the corresponding risk adjusted market return (Top).

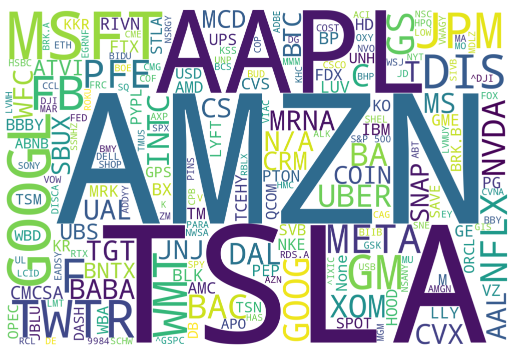

Figure 8: The most frequent mentioned stock tickers in the news.

factors as discussed in Section 4.4. Figure 11 illustrates an example of how a hybrid agent generates an analysis report from raw news input without iterative refinement. The analysis report is generated directly using the Top-5 relevant items from the memory with the following prompt:

> You are a helpful assistant designed to analyze the business news to assist portfolio management. Now, read this latest news and summarize it in one single paragraph, preserving data, date-time of the events, and key information, and include new insights for investment using the recommended relevant information:
>
> {news}

The architecture of the agent for the memory baseline mirrors that of the hybrid baseline. In the naive baseline, external memory is omitted, and the analysis report is produced directly from the refined news. The pricing network for the hybrid agent is identical to our method depicted in Figure 3, while the memory baseline omits the middle branch of manual factors. The naive baseline additionally removes the asset embedding branch.

E Prompts

In this section, we will present the prompts utilized by the agent, covering the refinement of the raw news input, the iterative refinement of the analysis report, the initial macroeconomic note, and the updating of notes.

E.1 News refinement

This refinement of the raw news input discussed in Section 3.1 is achieved through the following prompt:

> You are a helpful assistant designed to analyze business news. You need to use brief language to describe key information and preserve key data in the news. Now, analyze the following news: {input}

E.2 Iterative analysis

In the first iteration, the analysis begins with the following prompt:

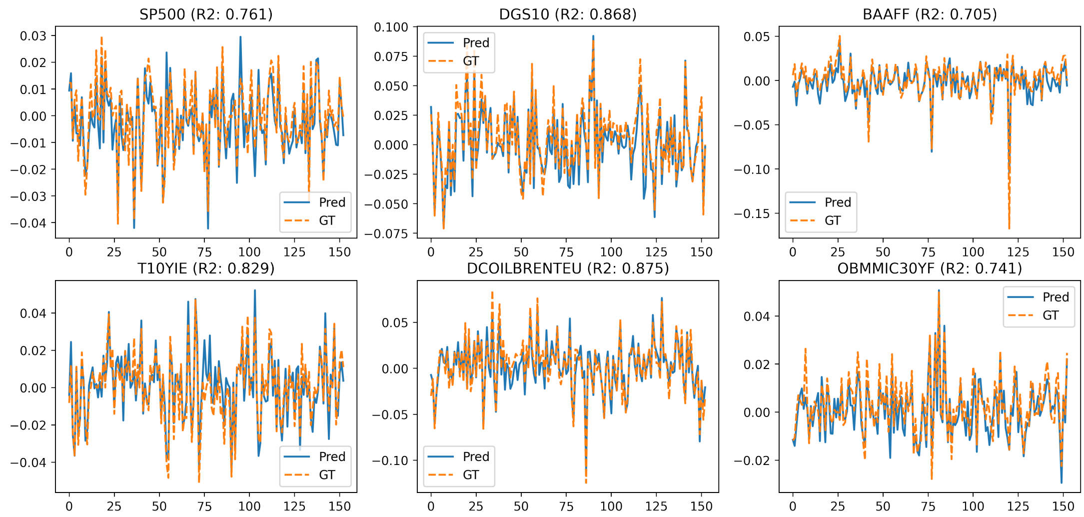

Figure 9: Use news features as the predictor to predict daily percentage change of the economic indicators.

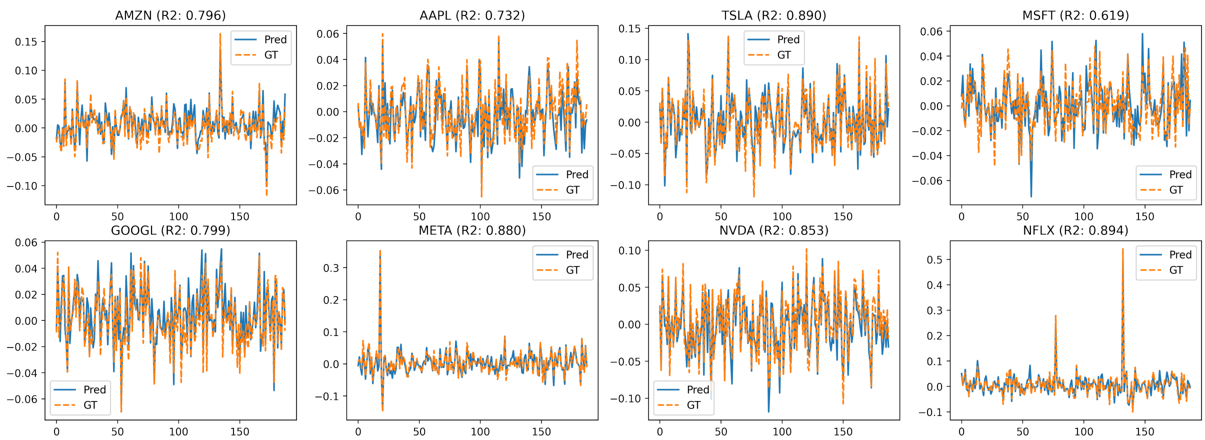

Figure 10: Predict the price movement of stocks in focus using augmented news features as predictors.

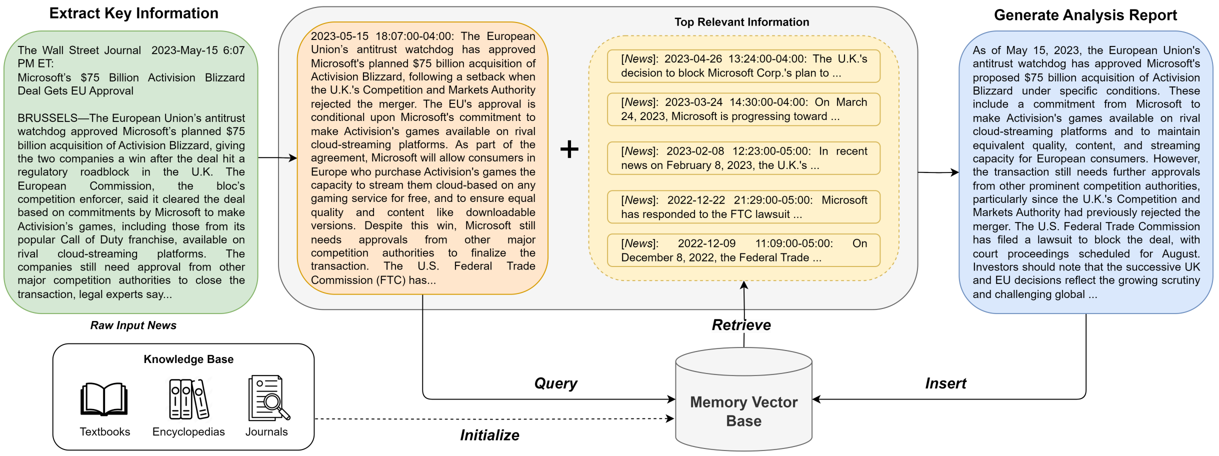

Figure 11: Example of the Hybrid agent baseline analyze a raw news without iterative refinement of analysis report as well as the macroeconomic and market trend notes.

You are a helpful assistant designed to analyze the business news to assist portfolio management.

> You will help me analyze this latest news from The Wall Street Journal and provide an analysis report, then I will search the relevant news or articles from the knowledge base based on your analysis report to help you refine it iteratively in multiple rounds. Let's start with this latest news, provide your analysis report, and I will help you refine it with the relevant information later, if you think this news is completely not helpful for investment now or future, call skip function to skip it, do not skip it if it may contain helpful information to future investment:
>
> {inputs}
>
> Here is a summary of the macroeconomics by today and the investment notes:
>
> {macro}

After the first iteration, the agent will be prompted as follows to continue the analysis:

> Based on your current analysis report, I found those potentially relevant news and excerpts from the knowledge base, please refine your analysis report with this information:
>
> {inputs}

In the last iteration, the agent will end the analysis with this prompt:

> Based on your current analysis report, I found those potentially relevant news and excerpts from the knowledge base, now finish your analysis report with them:
>
> {inputs}

E.3 Update of note

The macroeconomic and market trend note is updated with the new analysis report by the following prompt:

> Here is the current summary of the macroeconomic landscape and investment notes as of {date}:
>
> {macro}
>
> Now, given the latest news and the analysis report, update the macroeconomic summary with the new insights and impacts from the news. Include any relevant information that could influence the global economic outlook, such as geopolitical events, policy changes, or economic indicators. You should also take note of any important notes about investment trend and chances. Here are the latest news and the analysis report:
>
> {news}
>
> Now, update the macroeconomic summary with the new insights and impacts from the news as well as the investment notes.

E.4 Initial Macroeconomic Summary

We use the LLM to summary the macroeconomic status before the beginning time of the dataset, the summarization generated by GPT-3.5-Turbo-1106 is as follows:

> By September 2021, the global macroeconomic landscape was heavily influenced by the ongoing impacts of the COVID-19 pandemic. Many countries were in various stages of recovery, grappling with challenges such as disrupted supply chains, inflationary pressures, and shifts in employment patterns. Key points include: 1. **Economic Recovery**: Different regions experienced uneven recovery, with some economies bouncing back faster due to successful vaccination campaigns and substantial fiscal stimuli. For instance, the U.S. and China showed signs of robust economic rebound, whereas many European countries were still struggling with economic output below pre-pandemic levels. 2. **Inflation Concerns**: Rising inflation became a significant concern in many countries, partly due to supply chain disruptions and increased demand as economies reopened. This led to higher prices for commodities, goods, and services. 3. **Monetary Policy**: Central banks, including the U.S. Federal Reserve and the European Central Bank, maintained accommodative monetary policies, with low interest rates to support economic growth. However, there was growing discourse about when and how to start tapering these measures. 4. **Employment Fluctuations**: While some sectors and countries saw a rapid recovery in employment levels, others faced ongoing job losses, highlighting the pandemic's uneven impact across different industries. 5. **Supply Chain Disruptions**: Global supply chains were strained, impacting everything from consumer electronics to automobile manufacturing, leading to shortages and delays. 6. **Shifts in Consumer Behavior**: The pandemic accelerated trends like online shopping and remote working, reshaping economic activities and consumer behaviors in lasting ways. Overall, the state of global macroeconomics by September 2021 was defined by recovery efforts amidst ongoing challenges, with significant variability between different countries and regions.